DOCUMENT RESUME

ED 431 029 TM 029 862

AUTHOR	Andrade, Heidi Goodrich
TITLE	The Role of Instructional Rubrics and Self-Assessment in Learning To Write: A Smorgasbord of Findings.
SPONS AGENCY	Edna McConnell Clark Foundation, New York, NY.
PUB DATE	1999-04-21
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Grade 7; Grade 8; *Junior High School Students; Junior High Schools; *Self Evaluation (Individuals); Sex Differences; *Teaching Methods; Writing (Composition); *Writing Instruction
IDENTIFIERS	*Scoring Rubrics

ABSTRACT

Two studies were conducted to determine the impact of instructional rubrics on the development of students' writing skills and their understanding of the qualities of good writing. In the first study, 303 eighth graders were given an instructional rubric before writing 1 of 3 different types of essays. Results suggest that an instructional rubric can help students write better, but a more intensive intervention may be necessary to help all students perform at a higher level consistently. The second study examined the effects of instructional rubrics and guided self-assessment on students' writings and understandings of good writing. Students in 13 seventh- and eighth-grade classes wrote essays, but only those in the treatment group were given guided self-assessment techniques. Approximately 3 weeks later, 170 students completed questionnaires about the instruction. Self-assessment appeared to affect girls' writing favorably, but overall, self-assessment did not contribute more to students' overall knowledge of the qualities of good writing than did instructional rubrics alone. Three appendixes contain instructional rubrics for the three essay types. (Contains 3 figures, 5 tables, and 21 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 029

The Role of Instructional Rubrics and Self-Assessment in Learning to Write: A Smorgasbord of Findings

A paper presented at the annual meeting of the American Educational Research Association
April 21, 1999
Montreal, Canada

Heidi Goodrich Andrade
Harvard Graduate School of Education
Longfellow Hall—Project Zero
Cambridge, MA 02138

TM029862

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Heidi Goodrich
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Introduction

Scoring rubrics are among the most popular innovations in education (Goodrich, 1997a; Jensen, 1995; Ketter, 1997; Luft, 1997; Popham, 1997). However, little research on their design and their effectiveness has been undertaken. Moreover, few of the existing research and development efforts have focused on the ways in which rubrics can serve the purposes of learning and cognitive development as well as the demands of evaluation and accountability. The two studies described in this paper focus on the impact of instructional rubrics on the development of students' writing skills and their understandings of the qualities of good writing.

Theoretical framework

These studies draw on two areas of cognitive and educational research: authentic assessment and self-regulated learning. Perspectives on authentic assessment provide a guiding definition of assessment as an educational tool that serves the purposes of learning as well as the purposes of evaluation (Gardner, 1991; Goodrich, 1997b; Hawkins et al., 1993; Wiggins, 1989a, 1989b; Wolf & Pistone, 1991). In addition, the literature on authentic assessment provides guidance on the characteristics of effective assessment (see Goodrich, 1996a, for a review). These characteristics influenced the design of the studies reviewed below, which:

1. Articulated *clear criteria* for assessing writing,
2. Asked students to *assess their own work*,
3. Provided *opportunities for improvement* through revision, and
4. Was *sensitive to students' developmental stages*, referring to appropriate grade level standards.

The literature on self-regulated learning and feedback suggests that learning improves when feedback informs students of the need to monitor their learning and guides them in how to achieve learning objectives (Bangert-Drowns et al., 1991; Butler and Winne, 1995). My research is based on the hypothesis that students themselves can be the source of feedback, given the appropriate conditions and supports.

Taken together, the research on authentic assessment and on self-regulated learning point to the potential for instructional rubrics and self-assessment to support learning and skill development. In both of the studies reviewed below, these principles were made concrete by giving students instructional rubrics that describe good and poor writing (e.g., see Appendix A). I use the term "instructional rubrics" to refer to rubrics designed to support student learning and development in addition to serving as standards-referenced assessment tools. Instructional rubrics have several features that support student learning, including:

- they are written in language that students can understand;
- they refer to common weaknesses in students' work and indicate how such weaknesses can be avoided, and;
- they can be used by students to evaluate their works-in-progress and thereby guide revision and improvement.

Appendix A is an example of an instructional rubric I designed for use in this research. Like all of the rubrics I used, it draws on district, state and national standards as well as on feedback from colleagues and teachers. It articulates the criteria for the essay, describes gradations of quality from good to poor, and makes suggestions for avoiding typical writing pitfalls. The expectation in this research is that instructional rubrics, either alone or in combination with a formal process of self-assessment, will have significant effects on students' writing and learning.

Research Questions

This paper reports on two studies, each of which relied on instructional rubrics but were driven by different research questions. The research questions for each study were:

> Study 1: What effect does providing students with instructional rubrics have on students' writing and on their understandings of the qualities of good writing?
>
> Study 2: What effect does rubric-referenced self-assessment have on students' writing and on their understandings of the qualities of good writing?

Sample

This project was supported by the Edna McConnell Clark Foundation, which asked me to carry out the work in schools with which the foundation collaborates. As a result, the research was conducted in two middle schools in Southern California. One of the schools is located in a suburban community (School A), the other in an ethnically and linguistically diverse urban community (School B).

Measures

I collected data on two dependent variables for both studies: 1) students' scores on three essays written for this study, and 2) students' responses to a written questionnaire. The essays were scored by me and my assistants according to an adapted version of the rubrics used in the classroom intervention. Between 13% and 52% of the scores for each essay were tested for reliability.

The questionnaires consisted of one question: "When your teachers read your essays and papers, how do they decide whether your work is excellent (A) or very good (B)?" All students were asked to fill out the questionnaire approximately three weeks after they completed the final essay for this study.

I also collected data on several independent measures, including school attended, teacher, grade level, gender, ethnicity, previous performance in English as measured by ASAT scores and grades, and inclusion in ESL and special education classes.

Analysis

Multiple linear regression was used to understand the relationship between the treatment, the independent variables, and the essay scores. The main effect of each predictor and its interaction with the treatment condition were tested. Responses to the questionnaire were analyzed by noting the criteria to which students referred and comparing the treatment and control groups in terms of the number of references made to criteria contained in the rubrics used in this study.

Study 1—Instructional Rubrics

The first study spanned the 1996-97 school year and focused on the effects of instructional rubrics on eighth-grade students' writing and on their understandings of the qualities of good writing.

Procedure

Students in nine eighth-grade classes in the two participating schools were asked to write three different essays approximately one month apart: a persuasive essay, an autobiographical incident essay, and a historical fiction essay. Before writing a first draft of each essay, students in the treatment classes were given an instructional rubric. In one of the treatment classrooms, I introduced the rubric during one class period while the treatment teachers observed. The treatment teachers then introduced the rubric to their own classes while I observed. Students in the control classes were not given a rubric but were asked to write first and second drafts of the essays.

Results for Essay Scores

Table 1 lists the final regression models for each of the three essays. The parameter

Table 1
Final Regression Models for Essay Scores, Study 1

	Essay 1 n = 106	Essay 2 n = 37	Essay 3 n = 160
Intercept	1.57***	2.18**	1.62***
T_C	0.0009	0.49**	0.12
Grades	0.01***	-0.005	0.009*
ASAT	0.01*	0.01~	0.009~
Teacher	-0.10**		
School	0.30~	(N/A)	0.22*
Sex		-1.78~	0.51*
Grade*Sex		0.02~	
Ethnicity			0.20~
Trt*Sex			-0.43~
R^2%	25	40	19

~p<.10, *p<.05, **p<.01, ***p<.001

estimates and p-values for treatment condition (highlighted) reveal that there was a positive effect of the treatment on the second essay, the autobiographical incident, but not the first or third essays. Interestingly, the negative parameter estimate for the interaction between treatment and gender for Essay 3 (also highlighted) indicates that there was a negative effect of treatment on girls' scores on the autobiographical essay, but no effect for boys.

Essay 1. There was no measurable effect of the treatment on students' scores on Essay 1, the persuasive essay. The only statistically significant effects come from variables with traditionally robust predictive power: previous performance in English, teacher and school attended.

Essay 2. Because of implementation difficulties at School A during the writing of the second essay, the autobiographical incident, only essays from School B were scored. The results show that, controllin¸ for grades, ASAT scores, gender, and an interaction between grades and gender, treatment students are predicted to score, on average, almost half a point higher on a 4-point scale than control students. Figure 1 summarizes the effect of treatment on Essay 2 graphically.

-- insert Figure 1 here --

Essay 3. There appears to be a negative effect of treatment on girls' scores on Essay 3, the historical fiction essay. The interaction between treatment and gender approaches statistical significance at the .05 level, suggesting that the effect of treatment differs for girls and boys, controlling for grades, ASAT scores, school, and ethnicity. Since the main effect of treatment is not statistically significant, for boys there are no statistically significant differences in essay scores between the treatment and control groups ($t=.72$, $p=.47$), controlling for the other variables. For girls, however, the difference in predicted essay scores between the treatment and control groups approached statistical significance ($t=-1.74$, $p<.09$). The negative parameter estimate indicates that, on average, girls in the control group are predicted to have essay scores that are .31 points higher than girls in the treatment group, controlling for grades, ASAT scores, school and ethnicity. Moreover, the main effect of gender is statistically significant ($t=.2.22$, $p=.03$) which shows that, on average, girls in the control group are predicted to score .12 points higher than boys, controlling for grades, ASAT scores, and ethnicity. However, there was no statistically significant difference on essay scores between boys and girls in the treatment group ($t=.78$, $p<.43$), controlling for grades, ASAT scores, and ethnicity. Figure 2 represents this relationship graphically.

-- insert Figure 2 here --

Discussion of Essay Scores

Findings from the analysis of essay scores in Study 1 paint an uneven but intriguing pattern of results. In general, it appears that instructional rubrics can

help students write better but that a more intensive intervention may be necessary in order to help all students perform at higher levels consistently.

The lack of a treatment effect for the first essay may be due to the fact that it was many students' first exposure to a rubric. Only one of the eight teachers participating in this study had previously used rubrics. This is also a likely explanation for the fact that the teacher variable had an effect on scores on the first essay but not on the second or third essays. By the second and third essays, each of the teachers' classes had been exposed to rubrics. In addition, a power calculation suggested that this sample (n=106, control n=30) only had a power of 31% to detect a small effect of treatment even at the relaxed alpha level of .10. A larger sample size may or may not have detected an effect.

Findings from Essay 2 are more encouraging. The magnitude of the between-group differences for the second essay appears to be educationally as well as statistically meaningful. An average of a half-point difference on a 4-point scale is a 12.5% difference. This effect is all the more meaningful because of the minimal amount of classroom time taken by the intervention. Less than 40 minutes was spent on introducing and reviewing each rubric. Those 40 minutes may have translated into a 12.5% difference in students' scores.

The findings from the third essay stand in partial contrast to the findings from Essay 2. Essay 3 results indicate that instructional rubrics may actually create a detriment to the performance of girls but not boys. I suspect that girls in my sample may have responded more stridently to end-of-the-year pressures. Teachers at both schools reported that the third essay assignment came just as their students were attempting to meet portfolio and exhibition requirements for graduation. One teacher called it a "last ditch effort to complete their graduating exhibitions." This same teacher continued on to say, "Although the third essay would have been awesome to put in an exhibition, most kids were trying to take the easy way out (which was to revise something they already had rather than create something new). When push came to shove—finish exhibition and go to high school or finish the essay—high school won out." It may be that girls in the treatment group were more concerned about their graduation requirements or were more daunted by the demands of the third essay than were boys.

It is conceivable that the different results for each essay could also be explained in part by the fact that students were asked to write three different kinds of essays, and different kinds of writing require different kinds of skills. The historical fiction essay assignment was repeated during Study 2. I will make comparisons after discussing the results of that research.

Results of Questionnaire Analysis

Three of the four classes at School A filled out and returned the questionnaires, as did all five participating classes at School B. An analysis of students' responses to the questionnaire revealed striking differences between the treatment and control groups. As the following examples reveal, the control students tended to have a poorer understanding of how grades were determined:

> Well, they give us the assignment and *they know the qualifications* and if you have all of them you get an A and if you don't get any you get a F and so on (my emphasis).

Note that this student knows that the teacher has her standards or "qualifications" but he does not suggest that he knows what they are. The treatment students, on the other hand, tended to refer to rubrics, "rebeks" and "root braks" as grading guides and often listed criteria from the rubrics they had seen. For example:

> An A would consist of a lot of good expressions and big words. He/she also uses relevant and rich details and examples. The sentences are clear, they begin in different ways, some are longer than others, and no fragments. Has good grammar and spelling. A B would be like an A but not as much would be on the paper.

Many of the criteria referred to by this student were included in the rubrics he used during this study. I compared the criteria referred to by the control and treatment students. The responses from students in School A and School B were analyzed separately because the control students in School B had had previous exposure to rubrics used by their teacher. The control students at School A tended to mention fewer and more traditional criteria such as spelling, punctuation, and neatness. The treatment students, in contrast, tended to mention the same criteria to which the control group referred plus a variety of others, including criteria contained in the rubrics used in this study. Table 2 is a list of the criteria from the rubrics that were mentioned by treatment students at School A but not by control students. The numbers to the left represent the number of times each criterion was mentioned by students. School A control students did not refer to any of these eleven criteria, not even by chance.

Table 2
Criteria contained in rubrics and referenced by treatment students but not by control students at School A (n=74)

No. of references	Criterion
20	Word choice, e.g., "words give [the reader] a vivid picture in her mind"
8	Voice, reveals feelings and emotions
7	Interesting, not boring
3	Has accurate information
3	Provides details
2	Is descriptive
2	Uses proper paragraph format
2	Includes ideas, thoughts and opinions
2	Makes a point
2	Is well-organized, e.g., "has a beginning, middle and end"
1	Sentence structure

The results from School B are a little different because the control students were accustomed to using rubrics. Seven students in the control class referred to the use of rubrics in their responses, even though they were not given the rubrics used in this study. Nonetheless, small differences in the treatment and control groups at School B were found. Table 3 is a list of the criteria contained in the rubrics used in this study and mentioned by School B treatment students but not by control students.

Table 3
Criteria contained in rubrics and referenced by treatment students but not by control students at School B (n=122)

No. of references	Criterion
4	Word choice, "powerful words," "vividness"
4	Organization
3	Length, five paragraphs
3	Gives Details
2	Tells about action and events
2	Is easy to understand
2	Ideas and Content
1	Setting
1	The way the writing flows
1	Makes a point
1	Voice
1	Sentence fluency
1	Tells about lessons learned
1	Contains correct information

Discussion of Questionnaires

When compared to the responses of students in the control group, treatment students tended to refer to a greater variety of criteria for high quality writing. These differences suggest that the students who received instructional rubrics had more knowledge of what counts in good writing and of the criteria by which their essays were evaluated. It appears that instructional rubrics have the potential to broaden students' conception of good writing beyond the recognition of mechanics to include qualities such as word choice and voice and tone.

Study 2—Rubric-Referenced Self-Assessment

The second study took place during the 1997-98 school year and examined the effects of instructional rubrics and guided self-assessment on students' writing and understandings of good writing. This study involved thirteen seventh- and eighth-grade classes in the same two schools. Both the treatment and control groups wrote two essays. Students in all participating classes were given instructional rubrics, but only the treatment classes were engaged in a process of guided self-assessment.

Procedure

Students in each class were asked to write two different essays approximately one month apart: a historical fiction essay, and a response to literature. All classes were given identical instructional rubrics with the assignment of each essay, and their teachers briefly reviewed the assignment and the rubric.

After students had written a first draft of their essays (at least in theory), I conducted two self-assessment lessons. The first lesson guided students in using half of the rubric to evaluate their drafts in terms of the three most global criteria—ideas and content, organization, and paragraphs (see Appendix B). The treatment students were then asked to write a new draft and bring it to class for the second self-assessment lesson. During the second lesson I instructed them in using the second half of the rubric to look at the four finer grained criteria—voice and tone, word choice, sentence fluency, and conventions (see Appendix C).

The two self-assessment lessons focused on a formal process of guided self-assessment that I designed in collaboration with my participating teachers. We had students use markers to color code the criteria on the rubric and the evidence in their essays that showed that they met the criteria. A simple example comes from the historical fiction essay rubric, which includes a criterion requiring students to "bring the time and place in which the character lived alive." During class, I asked students to underline "time and place" in red on their rubrics, then underline the information they provided about the time and place of their story in red on their essay. If they could not find the information in their essay—and they were often shocked to discover they could not—I instructed them to write at the top of their papers a reminder to add the missing information when they wrote a second draft. This process was followed for all seven criteria on the rubrics. Control classes received copies of the rubrics but did not formally assess their own work in class.

As in Study 1, approximately three weeks after students completed the final essay for this study they were asked to respond to the one-question questionnaire. Some teachers did not have students complete the questionnaire, however, and others lost them. As a result, we had complete (treatment and control) data for the seventh-grade classes at School B, and two seventh- and two eighth-grade classes at School A. The total number of student questionnaires was 170 (85 treatment and 85 control).

Results for Essay Scores

Table 4 lists the final regression models for each of the essays. The parameter estimates and p-values for treatment condition (highlighted) reveal that there was no overall effect of the treatment on either the historical fiction essay or the response to literature essay. However, in the Essay 4 model, the parameter estimates for gender and for the interaction between treatment and gender (also highlighted) indicate that the effect of treatment differs by gender.

Table 4
Final Regression Models for Essay Scores, Study 2

	Essay 4 n=119	Essay 5 n=98
Intercept	.83*	1.11*
T_C	-.14	.07
ASAT	.15***	
Grades	.02***	.02**
Gender	-.47**	
Trt*Gender	.44*	
R^2 %	29.18	20.66

~p<.10. *p<.05. **p<.01. ***p<.001

Essay 4. Results of the analysis of students' scores on the historical fiction essay reveal an interaction between treatment condition and gender, after controlling for the other variables in the model. Figure 3 represents these results graphically. The red lines represent the effects for girls, the blue lines represent the effects for boys. The lines with triangles refer to the treatment condition. All four lines have an upward slope, indicating a positive relationship between ASAT scores and essay scores. There is also a positive, main effect of gender, such that boys consistently score higher than girls on Essay 4 (the solid lines are always above the dotted lines). In addition, there is a positive effect of treatment for girls—girls in the treatment group scored .31 points higher, on average, than girls in the control group (the red dotted line is above the blue dotted line). However, this effect is reversed for boys. Boys in the control group scored .14 points higher, on average, than boys in the treatment group.

Multiple regression analyses on the scores by gender revealed that the differences between treatment and control girls approach statistical significance (p=.08), but the differences for boys are not statistically significant (p=.39).

-- insert Figure 3 here --

Results of a multiple linear regression model using treatment, ASAT scores, grades (set to the sample mean), and gender to predict scores on Essay 4 (N=119)

Essay 5. Results for the response to literature essay revealed that there was no effect of treatment. The only significant main effect was the predictable effect of grades.

Discussion of Essay Scores Results

The analysis of essay scores from Study 2 indicate that self-assessment has no effect on students' writing, *and* that self-assessment can have a positive effect on girls' writing. I turned to the research on student response to feedback in order to explain the gender differences found in the fourth essay. In broad stroke, this

finding is consistent with research on sex differences in responsivity to feedback and in achievement motivation and learned helplessness. That body of research has generally shown that girls and boys differ both in their attributions of success and failure and in their response to evaluative feedback (Dweck & Bush, 1976; Dweck, Davidson, Nelson & Enna, 1978). However, the patterns found in Study 2 do not match those seen in Dweck's research. Briefly, research by Dweck and others (Deci & Ryan, 1980; Hollander & Marcia, 1970) has shown that girls are more likely than boys to be extrinsically motivated and to attribute failure to ability rather than to motivation or the agent of evaluation. As a result of these attributions, girls' performance following negative adult feedback tends to deteriorate more than boys' performance.

The findings from Study 2 are consistent with findings from an earlier study I conducted (Goodrich, 1996a) though, which showed that rubric-referenced self-assessment has a positive relationship with girls' metacognitive processing but a negative relationship with boys'. In combination, these studies suggest that self-generated feedback has a different effect than negative adult feedback on girls' performance. Some interesting contradictions in the research literature indicate that this finding may not be peculiar to my research. A study by Roberts and Nolen-Hoeksema (1989) found no evidence that women's greater responsivity to evaluative feedback led to performance decrements, suggesting that women's maladaptive responsivity to feedback is not absolute. Also of interest are earlier studies by Bronfenbrenner (1967, 1970), which found that when peers instead of adults delivered failure feedback, the pattern of attribution and response reversed: Boys attributed the failure to a lack of ability and showed impaired problem solving while girls more often viewed the peer feedback as indicative of effort and showed improved performance.

Noting that the more traditional finding of greater helplessness among girls was evident only when the evaluators were adults, Dweck et al. (1978) have taken these findings to mean "that boys and girls have not learned one meaning for failure and one response to it. Rather, they have learned to interpret and respond differently to feedback from different agents" (p. 269). This seems a reasonable conclusion to draw, and relevant to the gender differences found in this study. These studies did not allow me to examine students' attributions of success or failure, however, so this explanation of the differences between boys and girls is entirely speculative. The different ways in which boys and girls respond to self-assessment need to be better understood.

One final note about the results of Study 2: in the discussion of the results of Study 1, I speculated that some of the inconsistencies in the findings for the three essays might be explained by the simple fact that students were asked to write different kinds of essays each time. In order to investigate the validity of this explanation, I compared students' performance on the historical fiction essays written during each study. Such a comparison is of limited value because the interventions were different: for the first study the students just received a rubric and for the second they formally assessed their own work. Nonetheless, telltale patterns could be revealing if they exist—but they don't. In fact, the effects oppose one another. The Study 1 historical fiction essay treatment had a negative effect on girls and no

effect on boys. In Study 2, the treatment had a positive effect on girls and no effect on boys. I cannot definitively conclude that the kind of essay written has no effect on students' performance, but these findings do cast doubt on that argument.

Results for Questionnaires

There were remarkably few differences between the treatment and control students' responses to the questionnaire. Students in each group tended to refer to the same criteria and, when differences did exist, only between 1 and 5 students made mention of the criterion in question. I also looked at the number of references to rubrics, fairness, effort and "I don't knows," or a lack of knowledge about how a grade is determined by a teacher. The only apparent difference was in the number of students who considered their teacher's grading habits unfair. The control group (n=85) complained of unfairness 9 times, compared to 0 such complaints from the treatment group (n-85).

Discussion of Questionnaire Results

The analysis of students' responses to the questionnaire suggests that self-assessment did not contribute more to students' overall knowledge of the qualities of good writing than did the instructional rubrics alone. The questionnaire data suggest that self-assessment may decrease students' perceptions of unfairness in their teachers' grading practices, but not that it actually increases students' perceptions of fairness.

Conclusion

The analyses of the questionnaires from Study 1 indicate that instructional rubrics support the development of more sophisticated understandings of the qualities of good writing. Study 2 questionnaires indicate that self-assessment does not contribute much beyond what instructional rubrics contribute in terms of students' understandings of the qualities of good writing.

The results of the analyses of the essay scores are less straightforward. Table 5 summarizes the direction of the effects of the interventions on each essay, separated by boys and girls. The symbols in parentheses represent whether the treatment group performed better (+) or worse (-) than the control group when the differences did not reach statistical significance. The results from Study 1 suggest that it is possible that instructional rubrics support the development of students' writing skills and understandings of the qualities of good writing over time. Positive effects on writing are certainly not a given, however, and the effect of rubrics on girls' performance in particular needs further investigation. The results of Study 2 suggest that rubric-referenced self-assessment can have a positive effect on girls' writing but either a neutral or perhaps even a negative effect on boys' writing. Perhaps the safest conclusion to draw from this smorgasbord of findings is that *something* is happening. More qualitative and quantitative research is needed if the promises and pitfalls of instructional rubrics and self-assessment are to be understood and applied appropriately.

Table 5

The direction of the effects of the treatment on each essay, separated by boys and girls

	Study 1	Study 2			
Essay	1	2	3	4	5
Boys	0	+	0(+)	0(-)	0
Girls	0	+	-	+	0

References

Bangert-Drowns, R., Kulik, C., Kulik, J., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213-238.

Bronfenbrenner, U. (1967). Response to pressure from peers versus adults among Soviety and American school children. *International Journal of Psychology, 2*, 199-207.

Bronfenbrenner, U. (1970). Reactions to social pressure form adults versus peers among Soviet day school and boarding school pupils in the perspective of an American sample. *Journal of Personality and Social Psychology, 15* , 179-189.

Butler, D. & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245-281.

Deci, E., & Ryan, R. (1980). The empirical exploration of intrinsic motivational processes. In L. Berkowitz, (Ed.), *Advances in experimental social psychology*. New York: Academic Press.

Dweck, C., & Bush, E. (1976). Sex differences in learned helplessness: 1. Differential debilitation with peer and adult evaluators. *Developmental Psychology, 12* 147-156.

Dweck, C., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: 2. Contingencies of evaluative feedback in the classroom and 3. An experimental analysis. *Developmental Psychology, 14*(3), 268-276.

Gardner, H. (1991). Assessment in context: The alternative to standardized testing. In B. R. Gifford and M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction*. Boston: Kluwer.

Goodrich, H. (1997b). Thinking-centered assessment. In S. Veenema, L. Hetland, & K. Chalfen (Eds.), *The Project Zero classroom: new approaches to thinking and understanding*. Cambridge, MA: Project Zero, Harvard Graduate School of Education.

Goodrich, H. (1997a). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Goodrich, H. (1996a). Student self-assessment: At the intersection of metacognition and authentic assessment. Doctoral dissertation. Harvard University. Cambridge, MA

Hawkins, J., Frederikson, J., Collins, A., Bennet, D., & Collins, E. (1993). Assessment and technology: New ways to assess student learning. *Technology and Education, 36*(5).

Hollander, E. & Marcia, J. (1970). Parental determinants of peer orientation and self-orientation among preadoloscents. *Developmental Psychology, 2*, 292-302.

Jensen, K. (1995). Effective rubric design: Making the most of this powerful assessment tool. *Science Teacher, 62*(5), 34-37.

Ketter, J. (1997). Using rubrics and holistic scoring of writing. In Tchudi, S. (Ed.) *Alternative to grading student writing.* Urbana, IL: National Council of Teachers of English.

Luft, J. (1997). Design your own rubric. *Science Scope, 20*(5), 25-27.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership, 55*(2), 72-75.

Roberts, T., & Nolen-Hoeksema, S. (1989). Sex differences in reactions to evaluative feedback. *Sex Roles, 21*(11/12), 725-746.

Wiggins, G. (1989b). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41-47.

Wiggins, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wolf, D., & Pistone, N. (1991). *Taking full measure: Rethinking assessment through the arts.* New York: College Board Publications.

Appendix A. Instructional Rubric for Autobiographical Incident Essay

	5	4	3	2	1	N
Organization	My story has a strong lead that develops readers' interest, a developed middle that builds tension, and a satisfying ending that provides closure, all in an order that flows like water.	My story has either a strong lead, a developed middle or a satisfying ending but not all three. Maybe the middle drags on too long or the ending is a bit abrupt.	My story moves through the beginning, middle and end in a logical order. It takes the reader on a walk but on a sidewalk, not a high wire.	My organization is rough but workable. My story may get off topic once or twice.	My story is aimless or disorganized. It lacks direction.	I didn't write enough to say one way or the other.
Paragraphs	I indent the beginnings of all paragraphs and have one topic per paragraph. I wrote more than 5 paragraphs.	I indent the beginnings of all paragraphs, have one topic/paragraph, and I wrote 5 paragraphs.	Some of my paragraphs are too long or not indented. I wrote at least 5 paragraphs.	I have several problems with paragraphs and/or I wrote less than 5 paragraphs.	I use incorrect paragraph format and/or I wrote less than 5 paragraphs.	I didn't write enough to judge.
Action	My story gives details about one exciting, funny, sad or unusual event and reveals why it was important to me.	I tell about one specific event in detail but it isn't clear why it was important to me.	My story has one main event but also includes less important events that don't help readers understand what's important to me.	I focus on more than one event, none of which have enough detail to give the story a clear focus.	My story has no focus and is probably confusing to a reader.	I didn't write enough to judge.
Scene	I paint a mental picture for my readers, vividly setting the scene by describing important sights, sounds, smells, and/or tastes.	I describe the central scene(s) in detail, but not vividly.	I describe the scene at some point but some scenes are not described well.	I use only 1 or 2 descriptive words, only describe relatively unimportant scenes, or give irrelevant details.	I do not describe the setting of the journey.	I didn't write enough to judge.
Cast of characters	I create complex characters by showing them in action, describing how they look & act, by using dialogue and letting the reader "overhear" their inner thoughts.	I create characters by describing who they are, what they look like, gestures, expressions, and using relevant dialogue.	I tell who is in the story and their names and ages but do not show how characters behave and feel. I use little or only irrelevant dialogue.	I only vaguely refer to characters (e.g., I talk about "my brother" but never say his name, how he acts, etc.). I use no dialogue.	I leave significant characters out (e.g., my father, who took us on the trip I write about), and do not use dialogue.	I didn't write enough to judge.
The point	My paper reveals a profound insight gained from this trip. The lesson learned draws on a theme found throughout the essay.	I reveal insights gained from the trip, but they may be just tacked on at the end.	I describe relevant feelings or ideas, but I don't have a central insight or lesson learned.	I describe a few feelings or ideas but they aren't well connected to the story.	I don't share my feelings, insights or lesson learned. My essay seems to have no point.	I didn't write enough to judge.
Conventions	I use first person form, and correct sentence structure, grammar, punctuation and spelling.	My mechanics are good. Errors may be from taking risks, trying to say things in new or unusual ways.	I generally use the correct sentence structure, grammar, punctuation and spelling.	I make frequent errors which are distracting but do not interfere with meaning.	Numerous problems with grammar, spelling, etc. make my story hard to read	I better get busy writing!

H. Goodrich Andrade, Harvard Project Zero, 1998

BEST COPY AVAILABLE

Appendix B. Instructional Rubric for the Historical Fiction Essay, Part One

Criteria	5	4	3	2	1	N
Ideas and Content	My paper brings the time and place in which my character lived alive; vividly describes her/his experiences and values; refers to historically accurate facts and details.	My paper reveals the time and place my character lived; describes a day in her/his life; most or all details are historically accurate.	The time & place my character lived is clear, but his/her experiences are more like a list than a letter or diary entry; some details may be historically inaccurate.	I tell the reader when and where my story is set but make no attempt to include historically accurate facts or details.	The setting of my story is murky, and the characters' experiences and/or values are often historically inaccurate.	I didn't write enough to judge my own ideas and content.
Organization	My letter/diary has a strong lead, a developed middle, and a satisfying ending, all in an order that makes sense, flows, and hangs together.	I have either a strong lead, developed middle or satisfying ending but not all three.	I have a beginning, middle and end in logical order but without flair.	My organization is rough but workable. My writing may drag its feet then race ahead; my ending may stop suddenly.	My writing is aimless and disorganized; it is probably confusing to a reader.	I didn't write enough to judge.
Paragraphs	I indent the beginnings of all paragraphs & have one topic per paragraph. I wrote more than 5 paragraphs.	I indent the beginnings of all paragraphs, have one topic per paragraph, and I wrote 5 paragraphs.	Some of my paragraphs are too long or not indented. I wrote at least 5 paragraphs.	I have several problems with paragraphs and/or I wrote less than 5 paragraphs.	I use incorrect paragraph format and/or I wrote less than 5 paragraphs.	I didn't write enough to judge.

H. Goodrich Andrade, Harvard Project Zero, 1998

Appendix C. Instructional Rubric for the Historical Fiction Essay, Part Two

Criteria	5	4	3	2	1	N
Voice and Tone	I use 1st person. My voice is inviting and sounds like a real person. My paper has personality & shows how my character thinks and feels.	I use first person. I sound like I care about the topic. My writing voice is engaging but may fade in and out.	My tone is OK but my paper could have been written by anyone. I need to reveal more about how I think and feel about the topic.	My writing is bland, mechanical or pretentious. It sounds like I have not found my own way to say things. I may have used 2nd or 3rd person.	My writing is too formal or inappropriately informal. There may be no hints of a real person in it. It may sound like I don't like the topic. I may have used 2nd or 3rd person.	I didn't write enough to judge.
Word choice	The words I use are striking but natural, e.g., "wondered" instead of "thought." I use powerful verbs & historically accurate words, phrases and slang from the period.	My paper has some fine word choices and generally good language. Some parts may be routine.	The words I use are acceptable but ordinary. I should try to use more expressive words.	My word choice is uninspired, colorless, and dull or sounds like I am trying too hard to impress. Some words may be used incorrectly.	The same words are repeated over and over and over and over. Some words may be confusing to a reader.	I had better get busy....
Sentence Fluency	My sentences are clear; they begin in different ways; some are longer than others; no fragments; my paper is a delight to read out loud.	My sentences are mostly well constructed, some minor errors. My essay marches along but doesn't *dance*.	I have a few problems with fragments or awkward sentences that I should fix.	My sentences are often awkward or mechanical; little variety in length; may have many sentences that begin with the same word.	My paper is tough to read because almost all of my sentences are incomplete, run-ons, and/or awkward.	I didn't write enough to judge.
Conventions	I use the correct grammar, capitals, spelling, and punctuation.	I made some errors, perhaps by taking risks and using interesting words or sentences.	My spelling is correct on common words. Some errors in grammar and punctuation. I need to check it again.	Frequent errors are distracting but do not interfere with the meaning of my paper.	Many errors in grammar, capitalization, spelling & punctuation make my paper hard to read.	I didn't write enough to judge.

H. Goodrich Andrade, Harvard Project Zero, 1998

Figure 1.
Relationship between treatment, essay scores, grades, ASAT scores and gender for Essay 2, autobiographical incident (n=37)

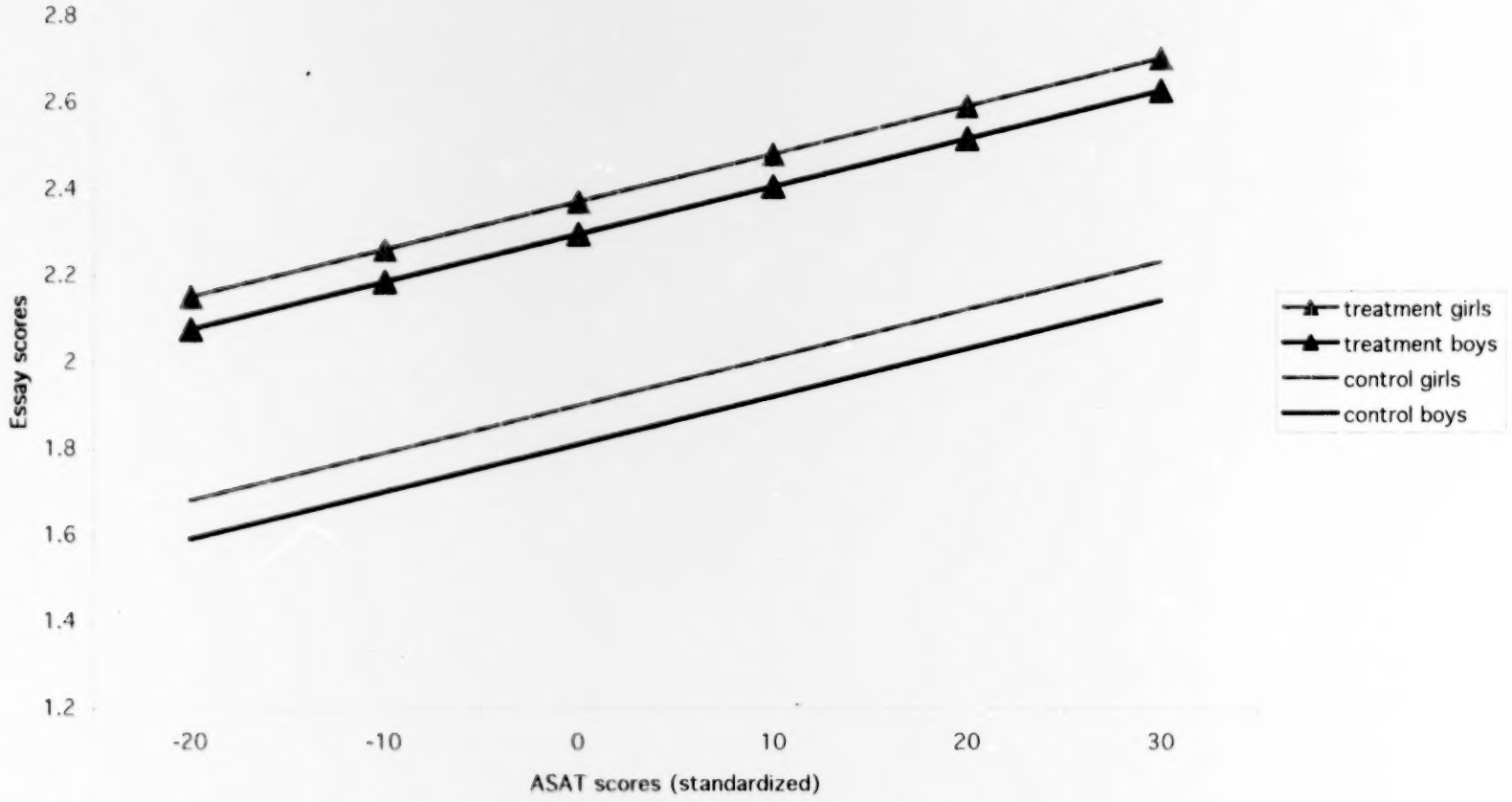

Figure 2.
Relationship between treatment, essay scores, grades and gender for Essay 3, historical fiction (n=160)

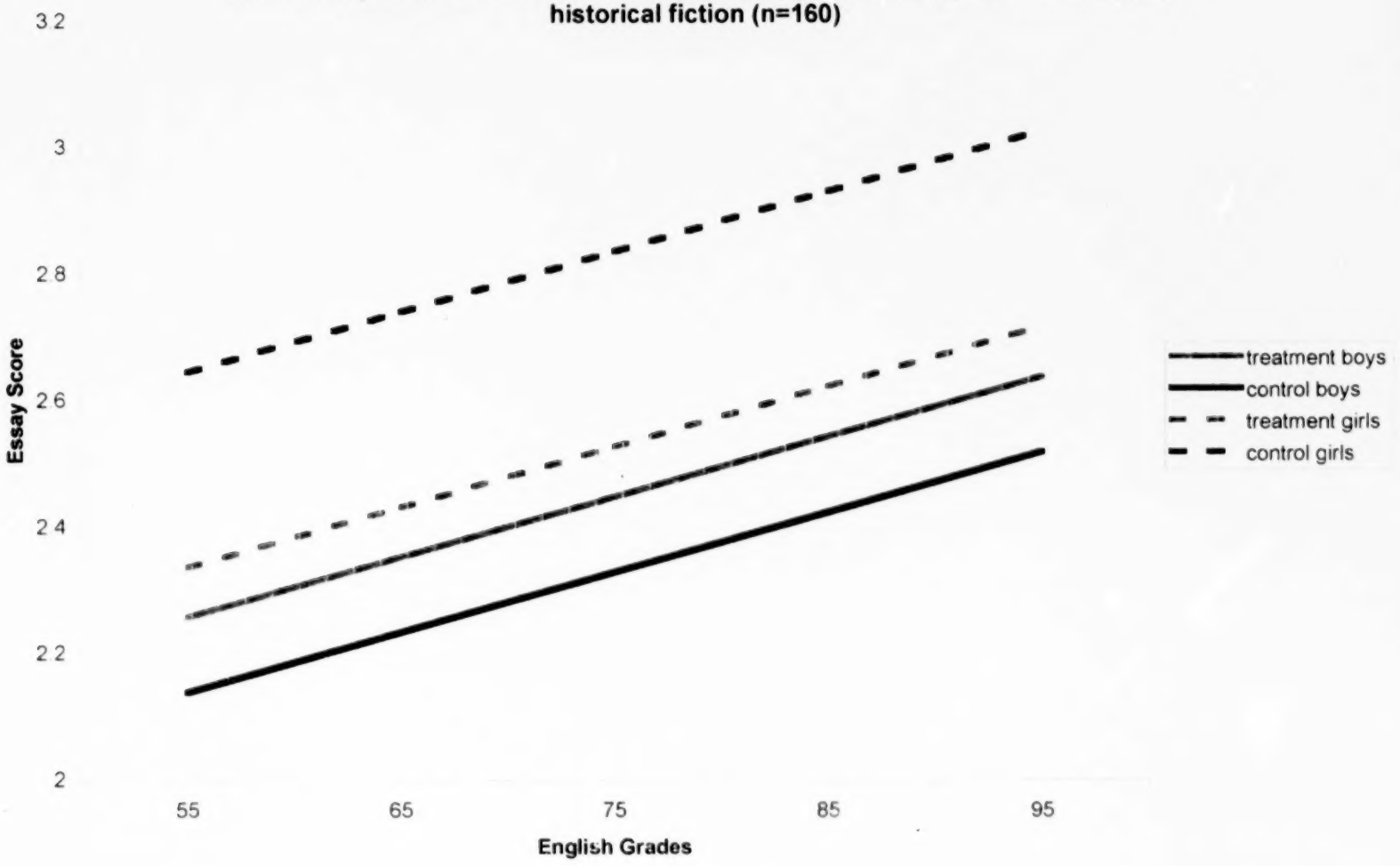

Figure 3.
Relationship between treatment, essay scores, ASAT scores, and gender for Essay 4, historical fiction (n=119)